enLLASD: An Ensemble Deep Learning Framework to Automate Derivation of Lower Limb Alignments for Skeletal Dysplasia

Peikai Chen, Xinlin Zhou, Haihua Cai, David J. H. Shih, Janus S. H. Wong, Yong Hu, *Senior Member, IEEE*, and Michael Kai-Tsun To

***Abstract*—Skeletal dysplasia (SD) is a group of rare, congenital skeletal disorders that affect millions worldwide. Patients often present with moderate to severe limb deformities, which need to be constantly monitored radiographically, in a process called alignment assessments. Currently, alignments are usually determined manually by physicians, with significant inter-rater variability and low efficiencies. Recent efforts to automate this often used individual deep learning models to detect bone contours or landmarks (e.g., joints). Due to data scarcity, case heterogeneity, and low bone mass, existing methods risk model overfitting. We propose enLLASD, an ensemble learning framework that integrates key-point detection and bone segmentation results from multiple member models, to enhance robustness. Predictions are aggregated using averaging, majority voting, and logistic stacking, while false positives in segmentation are suppressed through consistency with key-point predictions. A spline-based resampling method is used to fit the medial axes of the femur and tibia, enabling the computation of key alignment angles. We validated our framework on a dataset of 1416 full-length radiographs from both SD and non-SD individuals. Experimental results demonstrate that logistic stacking [intersection over union (IOU) for front femur: 0.9564, front tibia: 0.9468] outperforms individual models in segmentation accuracy and alignment angle estimation, particularly in cases with severe deformities. Our work highlights the potential of ensemble deep learning in automatic orthopedic radiography for SD.**

***Index Terms*—Deformity alignment, ensemble stacking, key-point detection, outline detection, skeletal dysplasia (SD), X-ray.**

Received 3 June 2025; accepted 27 June 2025. Date of publication 15 July 2025; date of current version 25 July 2025. This work was supported in part by Shenzhen Key Medicine Discipline Construction Fund under Grant SZXK077, in part by Shenzhen Clinical Research Center for Rare Diseases under Grant LCYSSQ20220823091402005, in part by the Sanming Project of Medicine in Shenzhen under Grant SZSM202311022, and in part by Theof Hong Kong-Shenzhen Hospital (HKU-SZH) Rare Diseases Seed Funding Project under Grant HKUSZH202404083. The work of Peikai Chen was supported by Shenzhen Peacock Plan under Grant 20210830100C. The Associate Editor coordinating the review process was Dr. Ting-Wei Wang. *(Corresponding authors: Michael Kai-Tsun To; Peikai Chen.)*

Peikai Chen, is with the Department of Orthopedics, The University of Hong Kong-Shenzhen Hospital (HKU-SZH), Shenzhen 518053, China, also with AIBD Lab, HKU-SZH, Shenzhen 518053, China, also with Shenzhen Clinical Research Center for Rare Diseases, Shenzhen 518053, China, and also with the School of Biomedical Sciences, HKU, Hong Kong, SAR, China (e-mail: pkchen@hku-szh.org).

Xinlin Zhou and Haihua Cai are with the Department of Orthopedics, The University of Hong Kong-Shenzhen Hospital (HKU-SZH), Shenzhen 518053, China, and also with AIBD Lab, HKU-SZH, Shenzhen 518053, China.

David J. H. Shih is with the School of Biomedical Sciences, HKU, Hong Kong, SAR, China.

Janus S. H. Wong is with the Department of Orthopedics and Traumatology, The University of Hong Kong (HKU), Hong Kong, SAR, China.

Yong Hu and Michael Kai-Tsun To are with the Department of Orthopedics, The University of Hong Kong-Shenzhen Hospital (HKU-SZH), Shenzhen 518053, China, also with AIBD Lab, HKU-SZH, Shenzhen 518053, China, also with Shenzhen Clinical Research Center for Rare Diseases, Shenzhen 518053, China, and also with the Department of Orthopedics and Traumatology, The University of Hong Kong (HKU), Hong Kong, SAR, China (e-mail: duqj@hku-szh.org).

Data is available on-line at github.com/HKUSZH/enLLASD.

Digital Object Identifier 10.1109/TIM.2025.3588983

I. Introduction

SKELETAL dysplasia (SD) is a large group of more than 700 diverse congenital conditions, over 90% of which are caused by genetic mutations related to extracellular matrix or its regulators, with major manifestations in the musculoskeletal system, including joints and bones [1]. Most SD are rare, with a combined incidence of only ~1/5000 live births, which still translates to millions affected worldwide. Some more commonly occurring SD include osteogenesis imperfecta (OI), chondrodysplasia (ACH), pseudoachondroplasia (PSACH), multiple epiphyseal dysplasia (MED), hypophosphatemic rickets, spinal muscular atrophy (SMA), etc. Despite their diversity and heterogeneity, many SD share clinical commonalities, including bone deformities, bone fragility, lower bone mass, and early onset age. Patients need to be constantly monitored with follow-up hospital visits, when X-ray radiographs are often taken and assessed (Fig. 1).

In clinical practice, such alignment assessments are typically performed manually by experienced clinicians using digital radiographs, which is a tedious, time-consuming process, and subject to significant inter-observer variability, especially in complex or severely deformed cases. Computerization of the process by vision-based approaches is desirable.

A. Related Works and Limitations

There have been many attempts to automate alignment measurements. Earlier works relied on traditional image processing and geometric modeling techniques [5], which often suffered from limited accuracy and poor generalization. Recent approaches based on deep learning and modern computer vision techniques have achieved much more accurate results [6]. A Harvard team applied a fully connected convolutional neural network (CNN) to 528 bilateral full-leg radiographs to estimate the hip–knee–ankle (HKA) angle, achieving a high correlation of 0.974 between predictions and ground truth [7]. Many such studies rely on identifying the landmarks, including hip, knee, and ankle centers. However, in real clinical practices, the bone contours are equally important in assessing skeletal deformities, especially the anatomical axes and center of rotation angle (CORA).

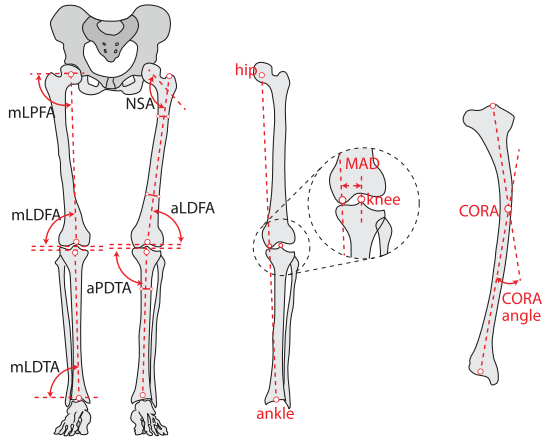


Fig. 1. Key alignment metrics in lower limb orthopedics in the frontal view. a: anatomical. m: mechanical. NSA: neck–shaft angle. mLDFA: mechanical lateral distal femoral angle. aLDFA: anatomical lateral distal femoral angle. mLPFA: mechanical lateral proximal femoral angle. mLDTA: mechanical lateral distal tibial angle. aMPTA: anatomical medial proximal tibial angle. MAD: mechanical anatomical deviations. CORA: center of rotation of angulation.

As such, some studies have attempted to bypass the key-point detection process entirely by focusing directly on bone region segmentation for structural recognition and parameter estimation. For instance, a study used a UNet model to segment the femoral head, knee, and ankle regions, and then calculated the centroids of these areas to estimate the HKA angle, thereby automating alignment evaluation without using explicit landmarks [7], [8]. In another study, researchers emphasized that bone contours are essential for capturing SD-specific structural deformities. In cases of enlarged epiphyses or extremely narrow diaphyseal shafts, key points alone are insufficient for describing such morphologies, and segmentation-based approaches are needed to calculate structural parameters such as the mid-diaphyseal angle [9], [10]. These works suggest an alternative whereby information of both bone contours (by segmentation) and landmarks can be fused to more faithfully infer bone morphology, thus deformity. It appears that such a framework remains lacking.

B. Problems and Challenges

Despite these progresses, automated alignment measurements still face significant challenges.

1) Due to their low-bone masses, radiographs from patients with SD often suffer from low signal-to-noise ratios (SNRs). Patients with SD are also diverse in clinical characteristics. Extreme cases of severe deformities are not uncommon. Robust strategies to handle severe and diverse cases are still missing.
2) Due to the rarity of SD, there are not many publicly available data. Thus, end-to-end state-of-the-art models that require large training data may risk overfitting. A modular approach may be more feasible.
3) Both the bone outlines and landmarks contain information that may help infer deformity. Proper strategies to fuse them to prune false positives or negatives, thus helping enhance robustness, remain elusive.
4) Most current approaches focus on estimating the HKA angle. Strategies for other important metrics, such as the neck–shaft angle (NSA), remain unaddressed.

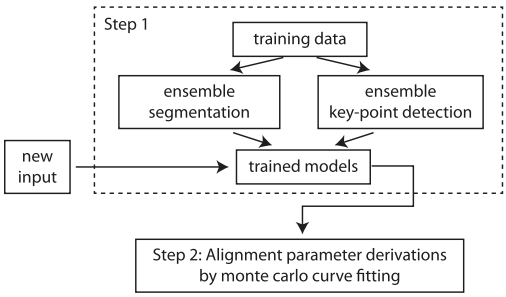


Fig. 2. Overall workflow of enLLASD. It consists of two major steps: ensemble flow to infer bone contours and landmarks, and parameter derivations.

C. Our Strategies

To address these problems, we propose enLLASD, an ensemble learning framework that integrates results from multiple individual models to enhance segmentation robustness. By jointly leveraging key-point detection and bone segmentation outputs, the framework achieves mutually reinforced predictions that suppress false detections and improve accuracy. Finally, spline-based resampling is applied to systematically derive a comprehensive set of clinically important alignment parameters. We declare that this study extends our preliminary findings presented at the 2024 IEEE CIVEMSA conference (paper ID: #1571014396) [3].

The main contributions of this work are as follows.

1) A scalable ensemble learning framework that allows the integration of any number of models trained on the same dataset. This design enhances robustness and generalization in the presence of complex or severely deformed skeletal structures.
2) A mutual filtering mechanism between key-point detection and bone segmentation is developed, wherein key points help pruning segmentation errors, and bone outlines assist in correcting misdetected key points. This mutual refinement leads to more accurate and stable alignment metrics.
3) A large full-length lower limb radiograph dataset containing SD and control cases is constructed, with annotated key points and bone contours. It supports further development and benchmarking in this field.

This article is organized as below. First, the current methods in alignment automation for SD are reviewed. Next, a set of solutions to address the problems are proposed and implemented, using an in-housely curated datasets of 1416 distinct X-ray images. Finally, the results are presented, both in terms of comparisons and representative images.

II. Methods

The enLLASD (Fig. 2) consists of two major steps: 1) an ensemble flow to infer skeletal morphology including bone contours and landmarks, and 2) the derivations of alignment parameters. We elaborate them below.

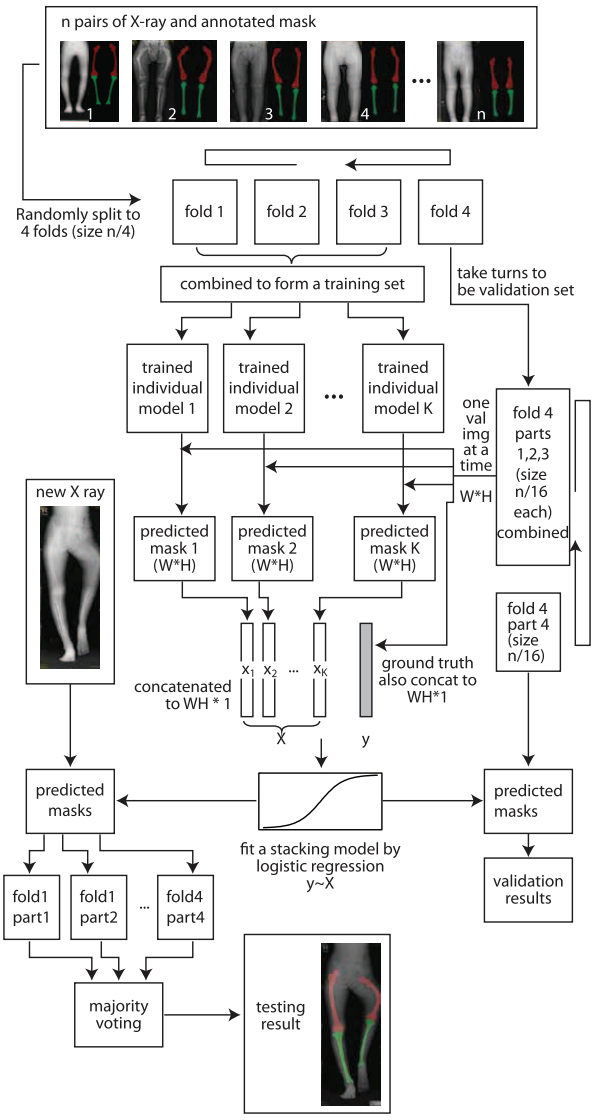


Fig. 3. Ensemble flow of enLLASD. The flow assumes K individual models and fourfold training. In each turn, the left-out fold was further split into four parts, with the parts taking turns in training the logistic stacking model. Given a new X-ray image, 4×4 predictions were obtained, and the final results were averaged by majority voting. Note the radiographic diversity of the patients in the input examples.

A. Step 1: Ensemble Flow to Infer Skeletal Morphology

Robustness is highly valued in medicine. Ensemble learning is known to reduce bias and/or variance, thus improving robustness [12]. We used ensemble approaches for bone landmark and contour detections, with details described below.

1) Ensemble Training Flow: As shown in Fig. 3, the training dataset is first randomly divided into k-fold (say 4), with each fold taking turns to be the validation set and the other three combined to form the training set. Multiple individual models were deployed and tuned for training to detect labeled landmarks or bone segment outlines on this set. Average, majority voting, or stacking approaches were used for either purpose to aggregate the individual results to make a final prediction.

2) Detecting Key Points: The key-point detection task aims to predict the 2-D coordinates of a set of anatomical landmarks, which help bone structure localization and subsequent analysis. Specifically, given a training dataset $\{\mathbf{X}_i \in \mathbb{Z}^{W_i \times H_i} | i = 1, 2, \ldots, N\}$ and supervised labels $\{(x_{i,j}, y_{i,j}) \in \mathbb{Z}^2 | i = 1, \ldots, N; j = 1, \ldots, M\}$, where N is the total number of training X-ray radiographs, W_i and H_i are the respective width and height of the ith image, and M the number of key points (a.k.a. landmarks) per image, and the objective is to train a model $f(\cdot)$ such that given a new input $\mathbf{X}$, it produces an estimate $f(\mathbf{X}) = \{(\hat{x}_j, \hat{y}_j)\}$ of the coordinates of the landmarks, with minimal total loss

$$J = \sum_{j=1}^{M} \text{loss}\left(\left(\hat{x}_j, \hat{y}_j\right), \left(x_j, y_j\right)\right) \tag{1}$$

where $\{(x_j, y_j)\}$ are the ground-truth coordinates for the new input.

In the ensemble learning framework for key-point detection, we first fuse the predictions from multiple models using arithmetic averaging to obtain more robust key-point location estimates

$$\left(\hat{x}_j^{en}, \hat{y}_j^{en}\right) = \frac{1}{K} \sum_{k=1}^{K} f_{k,j}(\mathbf{X}). \tag{2}$$

Theoretically, we expect this fusion strategy to reduce the overall prediction error. Given K models $f_k(\cdot)$, the objective is to find a set of weights w_k that minimize the ensemble loss J_{en} for key-point detection

$$J_{en} = \sum_{j=1}^{M} \text{loss}\left(\left(\hat{x}_j^{en}, \hat{y}_j^{en}\right), \left(x_j, y_j\right)\right) \tag{3}$$

with $(\hat{x}_j^{en}, \hat{y}_j^{en}) = \sum_{k=1}^{K} w_k \cdot f_{k,j}(\mathbf{X})$, and $f_{k,j}(\cdot)$ is the kth model's estimate on the jth key point.

In addition, when the bone segmentation results were pruned, we may calculate certain key points (e.g., the knee center, Fig. 4) from the segmentation alone and use this information to update the ensemble results, in a Bayesian–Kalman manner. Let (x_0, y_0) denote $(\hat{x}_j^{en}, \hat{y}_j^{en})$ for the knee, and

$$\begin{bmatrix} x \\ y \end{bmatrix}_{\text{prior}} = \begin{bmatrix} x_0 \\ y_0 \end{bmatrix}, \quad \mathbf{\Sigma}_{\text{prior}} = \begin{bmatrix} \sigma_{x0}^2 & 0 \\ 0 & \sigma_{y0}^2 \end{bmatrix} \tag{4}$$

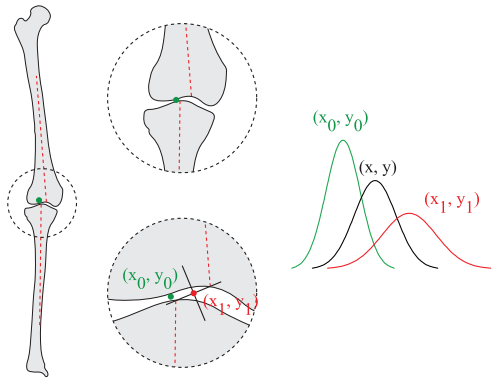


Fig. 4. Strategies to update key-point estimation for the knee. Green dot and green curve are the ensemble estimate, and red dot and red curve are the geometric knee center derived from the two bone segments. The three bell-shaped curves represent the Bayesian–Kalman approach to update the knee center estimate (black curve), with information from ensemble key-point result (green) and the geometric estimate (red).

where σ_{x0}^2 and σ_{y0}^2 can be estimated from $\sigma_{\hat{x}_j^{en}}^2$ and $\sigma_{\hat{y}_j^{en}}^2$, respectively. Let (x_1, y_1) denote the observed geometric knee center (red dot in Fig. 4) derived from the segmented bones

$$\begin{bmatrix} x \\ y \end{bmatrix}_{\text{observation}} = \begin{bmatrix} x_1 \\ y_1 \end{bmatrix}, \quad \mathbf{R} = \begin{bmatrix} \sigma_{x1}^2 & 0 \\ 0 & \sigma_{y1}^2 \end{bmatrix} \tag{5}$$

where σ_{x1}^2 and σ_{y1}^2 can be estimated from the "knee area," empirically set as a circle with diameter equal to the width of the distal end of the femur. This setting accounts for individual variability by adjusting the region of interest proportionally, approximates the actual joint contact area, and ensures coverage of key structures without including excessive unrelated tissues. The updated posterior (x, y) is then given by the following equation:

$$\begin{bmatrix} x \\ y \end{bmatrix}_{\text{posterior}} = \begin{bmatrix} x_0 \\ y_0 \end{bmatrix} + \mathbf{K}\left(\begin{bmatrix} x_1 \\ y_1 \end{bmatrix} - \begin{bmatrix} x_0 \\ y_0 \end{bmatrix}\right) \tag{6}$$

with $\mathbf{K} = \mathbf{\Sigma}_{\text{prior}}(\mathbf{\Sigma}_{\text{prior}} + \mathbf{R})^{-1}$. Effectively, the new coordinates are the variance-weighted mean of the ensemble-predicted and segmentation-derived positions.

3) Detecting Bone Contours: Detecting bone outlines is an image segmentation problem, and we adopt a semantic segmentation approach. Similar to the key-point task, a set of images $\mathbf{X}_i$ and corresponding masks $\mathbf{Y}_i \in 0, 1^{W_i \times H_i}$ are provided as supervision, where 0 denotes background (nonbone) and 1 denotes bone. This is formulated as a pixelwise binomial classification problem.

The model $f(\cdot)$ is trained to produce a probability matrix $f(\mathbf{X}) = \hat{\mathbf{Y}} \in [0, 1]^{W_i \times H_i}$ that minimizes the loss

$$J = \sum_{i,j} \text{loss}\left(\hat{Y}_{i,j}, Y_{i,j}\right). \tag{7}$$

In the ensemble case, we aim to improve robustness through multiple models' predictions. The ensemble loss is defined as follows:

$$J_{en} = \sum_{i,j} \text{loss}\left(\hat{Y}_{i,j}^{en}, Y_{i,j}\right) \tag{8}$$

where $\hat{Y}_{i,j}^{en} = g(f_k(\mathbf{X}))$, and $g(\cdot)$ is the ensemble function.

In each fold k, we use a stacking strategy. After predictions from all K models are obtained, a logistic regression is trained to combine them (see Fig. 3). The images in fold k are further divided into multiple parts (say 4) and take turns to train the logistic predictor. Specifically, suppose there are 100 images in fold k, 75 of them would be for training the stacking model $\mathbf{y} \sim 1/[1 + \exp(-\mathbf{X}\boldsymbol{\beta})]$, where $\boldsymbol{\beta} \in \mathbb{R}^K$, $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_K] \in \mathbb{Z}^{n \times K}$, and $\mathbf{y} \in \mathbb{Z}^n$. K is the number of individual models (say 6). The fit coefficients $\hat{\beta}$ can then be used to predict new testing inputs.

4) Fusing Bone Contours With Landmarks for Pruning: Due to the low bone density in patients with SD [13], their X-rays often exhibit extremely low SNRs, causing significant false positives or negatives even with ensemble models. To mitigate this, we use the anatomical priors from key-point detection to prune unreasonable regions in segmentation (Fig. 5). For example, tibia predictions above the knee or below the ankle can be removed. Similarly, femur predictions below the knee or above the femoral head can also be discarded. This spatial pruning significantly improves segmentation robustness.

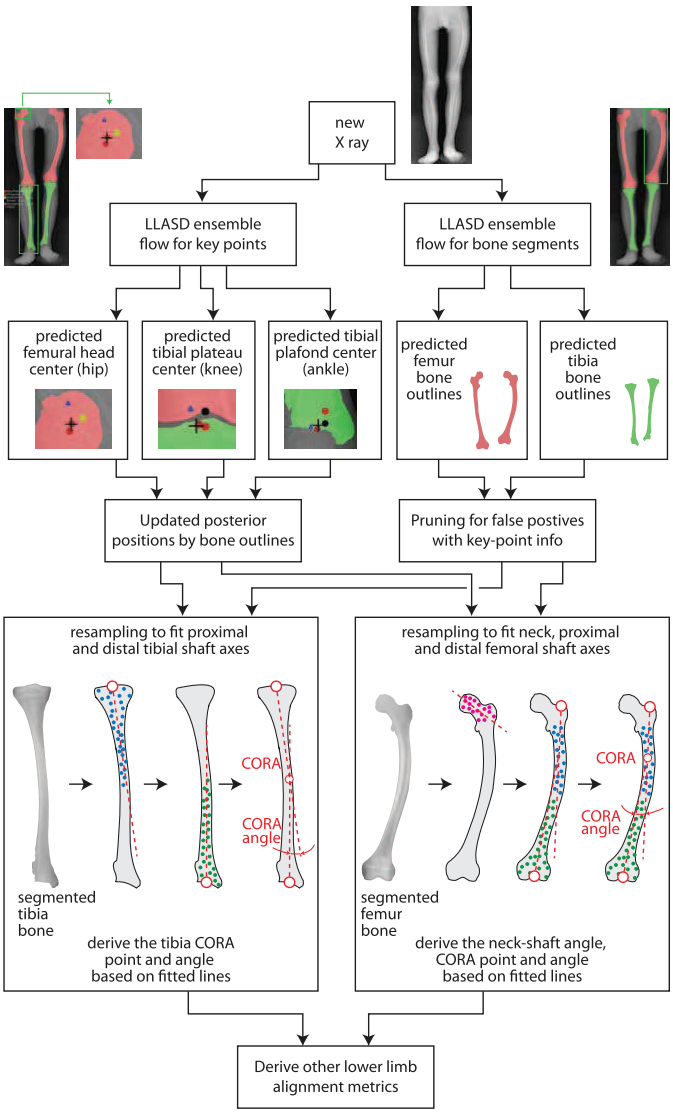


Fig. 5. Alignment derivation strategies of enLLASD. This represents the inference stage, using individual and ensemble models trained in the previous stage. Example images were shown alongside each step.

B. Step 2: Deriving Alignment Metrics

Upon completion of Step 1, for a given new input image $\mathbf{X}$, the next task of enLLASD is to derive the alignment metrics, such as CORA angle and NSA (Fig. 1).

For angular metrics such as CORA or NSA, which typically involve two slopes k_1 and k_2 of two lines, the angle in degrees is straightforward with simple geometry, given by: $\theta(k_1, k_2) = 180/\pi \cdot \arctan|(k_1 - k_2/1 + k_1 k_2)|$.

In clinical practice, when landmarks such as the femoral head center, knee center, and ankle center are known, anatomical axes (e.g., femoral neck, proximal femur shaft) can be manually constructed, and the corresponding angles are straightforward to derive. However, in automatic image processing settings, some of these anatomical landmarks are difficult to identify, particularly the greater trochanter tip for

femoral neck axis fitting, due to variability in local shape contours. Indeed, many common angular metrics (e.g., NSA) have been reported to be ill-defined even in manual annotation scenarios [11].

To address this, we adopt a stochastic axis estimation approach. Specifically, for each segment of interest (e.g., femoral neck, proximal/distal femur shaft, proximal/distal tibia shaft), a subset of pixels is sampled from the segmentation result and used to fit an axis. This process is made possible by the previously refined key-point positions and pruned segmentation masks.

Three methods were explored for line fitting: ordinary linear regression, RANSAC [14], and principal component analysis (PCA). Linear regression may produce inconsistent results when the slope is steep, due to its assumption that errors arise solely from the y-axis (or x-axis if inverted). In contrast, PCA estimates the principal axis of variance using both the x and y dimensions, which is more suitable for steep bone shaft structures.

In practice, we found that sampling thousands of points from the upper or lower 40%, 50%, or 60% regions of the tibia shaft produced stable medial axis slopes. Similarly, fitting the femur shaft was relatively easy. However, accurately estimating the femoral neck axis proved to be challenging due to its short and irregular structure. To tackle this, we first selected a small percentage (e.g., 10%) of the points closest to the femoral head center, performed a line fit, and then gradually increased the percentage. The process stopped once the newly included points aligned more with the femur shaft axis rather than the neck, ensuring that the final line captured the actual femoral neck direction.

III. Experiments and Results

A. Data and Labeling

Medical records of all the orthopedic patients between 3 and 55 years of age (mean 8.9 years, standard deviation 7.1 years; 60.3% male and 39.7% female) undergoing inpatient or outpatient treatments at our hospital between 2012 and 2023 were retrieved. Institutional review board (IRB) approvals were obtained with approval numbers: [2022]-191 and [2024]-305. Patients without lower limb radiographs were excluded. Patients with existing metal implants, including rodding and locking plates, were also excluded.

The labeling of the key points and bone outlines for the data was conducted on LabelMe (version 5.2.1) [15]. XZ and HC did the bulk of the labeling, and the results were reviewed by two experienced orthopedic surgeons (MT and JW). Specifically, the femoral head center (hip), tibial plateau center (knee), and tibial plafond center (ankle) of each radiograph were labeled manually. For outline labeling, we first used segment anything [16] to provide a rough outline, which drastically reduced the workload, followed manual fine-tuning adjustments. No distinction was made between left and right legs in bilateral images.

Consequently, 1416 distinct full-length lower limb X-ray images, including 440 bilateral (double-leg) anteroposterior (frontal view) radiographs and 976 lateral (side view) unilateral (single-leg) radiographs, were obtained. Images deemed unfit for annotation, including indistinguishable bone contours or anatomical landmarks, were excluded, leading to 432 and 368 frontal images with key-point and bone segment outline annotations, respectively, and 776 and 802 lateral images for key-point and bone segment outline annotations, respectively (Table I).

TABLE I
Numbers of Curated Distinct Full-Length Lower Limb Radiographs

	Front (anteroposterior)		Side (lateral)	
	non-SD	SD	non-SD	SD
Key-point	216	216	388	388
Segmentation	184	184	450	352

B. Model Selection and Deployment

We used six models for key-point detection ($K = 6$).

1) Yolo (V8) [17].
2) Stacked hourglass networks [18].
3) ShuffleNet [19].
4) HR-Net [20].
5) MobileNet (V2) [21].
6) RTMpose [22].

Of these, Yolo is end-to-end, with direct prediction of key-point coordinates, whereas the other five models involve first locating rectangular region before performing point prediction. Toward this end, the region proposal network Faster R-CNN model was adopted to locate the legs before those joint-detection models were trained [23].

Apart from YOLO, all other models were trained under the openmmlab environment [32] with default configuration files. Briefly, the mean squared error (mse) loss with weight equals to 1.0 was used. For key-point detection, a maximum of 250 epochs (amounting to 5000 iterations), with a batch size of 32 were used. The Adam optimizer with a learning rate of 5×10^{-4} was set. For key-point detection, the normalized distance percentage of correct key points (PCKs) metric was used.

For segmentation tasks, we also chose six member models ($K = 6$).

1) Fast S CNN [24].
2) DeepLabv3 [25].
3) KNet [26].
4) Pyramid scene parsing network (PSP-Net) [27].
5) UNet [28].
6) SegFormer [29].

For segmentation models' training, a maximum of 20 000 iteration were set. The weight decay optimizer AdamW with learning rate 6×10^{-5} and weight decay 0.0005 was used. Pixelwise cross entropy loss $-\sum_{s=1}^{2} y_s \log(\hat{y}_s)$, where $\hat{y}_s \in [0, 1]$ is the output probability for state s and $y_s \in \{0, 1\}$ is the truth, was used. Random resizing, cropping, and flipping were set to enforce data augmentation. The model with best intersection-over-union (IOU) metric was kept for validation.

In both the key-point and segmentation models, the front (anteroposterior) and side (lateral) images were trained and tested separately. A fold size of $F = 4$ was chosen. In either view group (front or side), the non-SD (controls) and SD (cases) data were combined, though, leading to a fold size

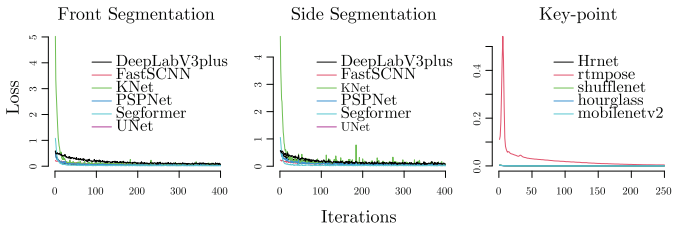


Fig. 6. Individual model training process.

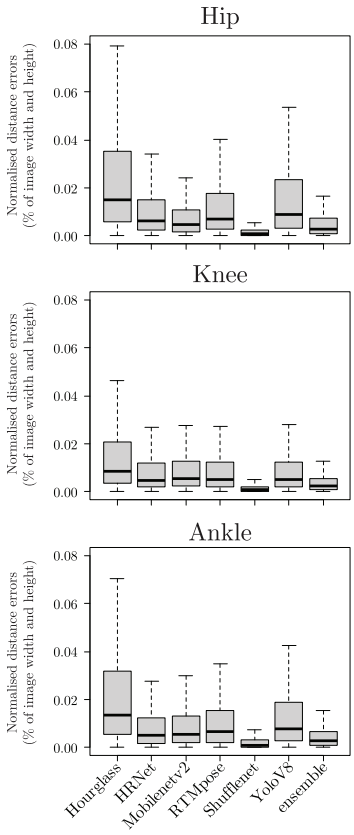


Fig. 7. Key-point prediction results. The vertical axes indicate normalized deviations, while the horizontal axes indicate models.

of 108 (= 216 × 2/4), 194 (= 388 × 2/4), 92 (= 184 × 2/4), and 200 (=(450 + 352)/4) for the key-point front, key-point side, segmentation-front, and segmentation-side groups, respectively.

In performing downsampling for shaft axis fitting, 5000 data points were uniformly sampled from each predicted bone segment mask, which usually numbers in tens of thousands. The process was repeated 100 times to assess variability, in a Monte Carlo manner.

The models were trained on an advanced workstation with 256GB RAM and two GPUs (NVIDIA GeForce RTX 3090), and a reasonable training convergence was achieved (Fig. 6).

Example testing radiographs were also tested with SOTA segmentation models, including segment anything model 2 (SAM2) [30] and MedSam [31], with manual prompts of region of interest, but without fine-tuning.

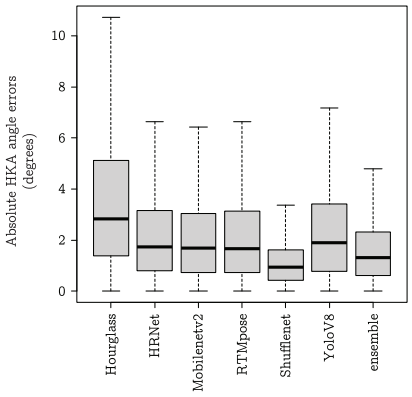


Fig. 8. HKA angle evaluation. The error was calculated based on absolute differences between the HKA angles derived from the labeled key points (ground truth) and those derived from the predicted key points.

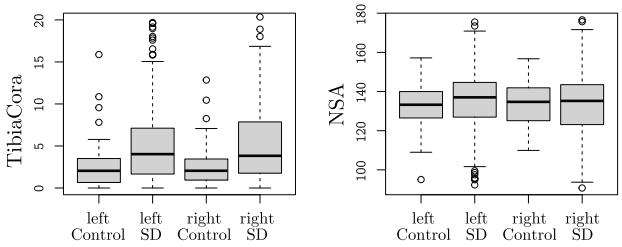


Fig. 9. Alignment estimation results. The vertical axes indicate estimate of alignment angles, including thigh CORA angle (left panel) and NSA (right panel), and the horizontal axes indicate the left and right legs of control or SD patient groups.

TABLE II
COMPARING SEGMENTATION RESULTS (IOUS)

	Femur	Tibia
Front		
stacking	**0.9564±0.007**	**0.9468±0.0066**
majority	0.9501±0.0075	0.933±0.0107
PSPNet	0.9342±0.0105	0.9148±0.0121
UNett	0.8064±0.02	0.8271±0.0346
KNet	0.9555±0.0052	0.9456±0.0072
FastSCNN	0.8552±0.0235	0.8342±0.0213
DeepLabV3plus	0.9495±0.0106	0.9377±0.0097
Segformer	0.8973±0.0251	0.8555±0.0263
Side		
stacking	0.9466±0.0108	0.9461±0.0081
majority	0.9012±0.024	0.9293±0.0125
PSPNet	0.8704±0.037	0.9037±0.0196
UNet	0.8153±0.0465	0.9017±0.0264
KNet	**0.9484±0.0102**	**0.9504±0.0044**
FastSCNN	0.8035±0.0462	0.8497±0.0278
DeepLabV3plus	0.8749±0.0425	0.9095±0.0282
Segformer	0.7735±0.049	0.7864±0.0373

± indicates sample standard deviation.
majority is when a pixel is predicted to be positive in > 3 models

C. Key-Point Detection Results

The predicted errors in terms of normalized Euclidean distances are shown in Fig. 7, with derived HKA and NSA shown in Figs. 8 and Fig. 9, respectively. It shows the advantages in using an ensemble approach. Although ShuffleNet outperforms other individual models in the current training, the ensemble result is comparable in achieving low prediction errors. Critically, the ensemble approach in enLLASD is expected to reduce the chance of spurious outliers, a highly undesirable situation in the medical diagnostic setting.

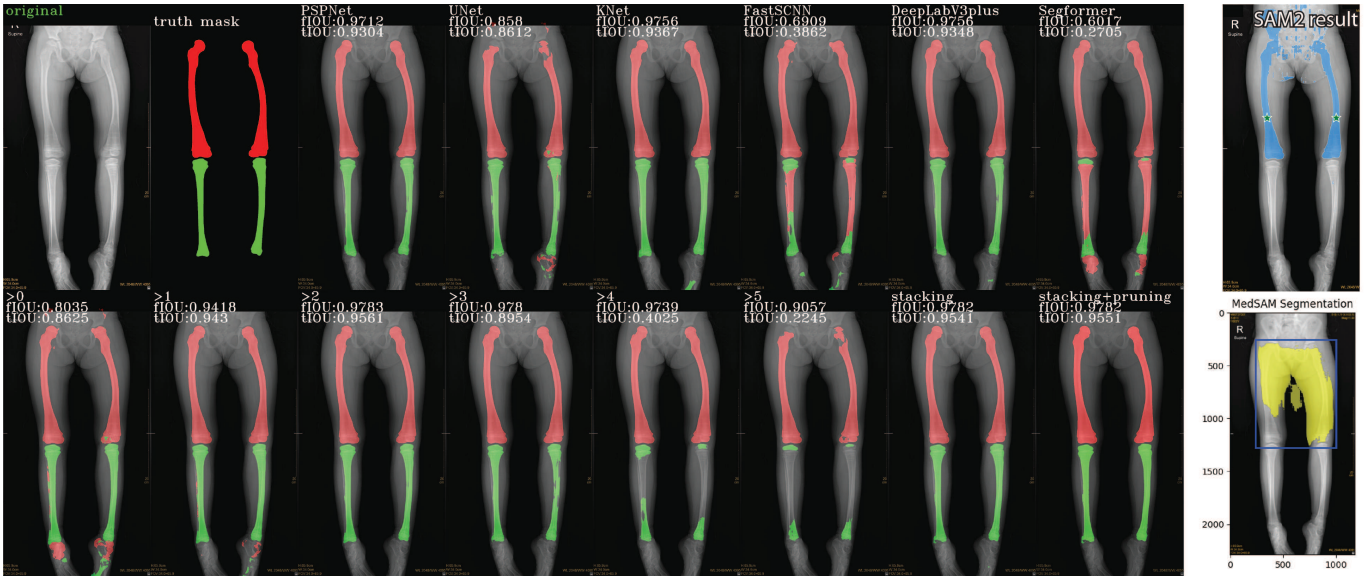


Fig. 10. Representative segmentation output of an SD example with low SNR X-ray. Stacking refers to ensemble logistic regression. ">0" to ">5" refers to threshold voting of the six individual model outputs in the first row. ">0" is the union, and ">5" is the intersection. ">3" is majority voting. fIOU and tIOU are the IOU with ground truths in the femur and tibia, respectively. "Stacking + pruning" refers to pruning of false positives after the stacking ensemble step. Results from (SAM2, stars representing prompts and blue area as segmented masks) and MedSam (blue boxes as prompt, yellow area as segmented masks) were also shown for comparison.

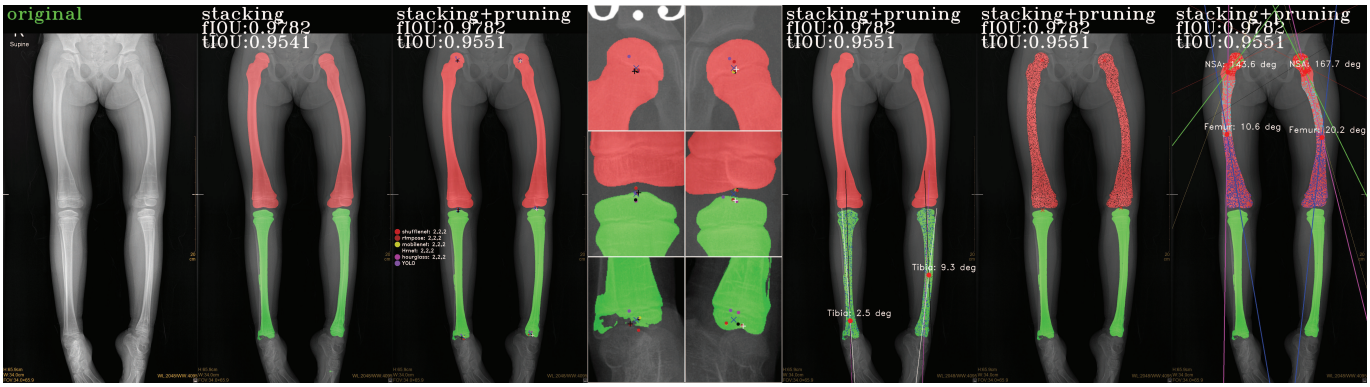


Fig. 11. Representative output of the alignment derivations in an SD example in the frontal view. First from left: original X-ray image. Second: identified ensemble masks for femur (red) and tibia (green). Third: postpruning segmentation results, with individual and ensemble prediction results for key points of the joints. Fourth: zoomed-in view (4×) of the six joints. Blue tilted crosses are ensemble results, and black and white crosses are human labeled landmarks for the left and right legs, respectively. Fifth: fitting tibia shaft axes for estimating CORA. Sixth and seventh: fitting femur shaft axes, neck axis for estimating NSA and femur CORA. Note that in the last three subplots, the color dots are downsampled data points for fitting the shaft axes. In addition, in the last subplot, the green line in the femoral neck is fit by PCA, and the black by RANSAC method.

D. Segmentation Results

Table II shows the comparisons between majority voting, logistic stacking, and other individual member models in terms of IOUs. For the front images, stacking method outperforms all other approaches. In the side images, KNet is the best, but the stacking results are also quite close and are much better than other individual models. The representative output in Fig. 10 shows that even for very dim X-rays, the detected masks still nicely capture the bone outlines, and that the pruning further improves the results. In comparison, the SAM2 and MedSam, even with manual human prompts on the femurs (stars and bounding boxes, right panel, Fig. 10), produced results that are unfit for any medical use. It is possible that fine-tuning of MedSam with our training dataset may substantially improve the results. We argue that in that case, it may be added as a member model (the seventh model) of our ensemble approach and help improve our results further.

E. Alignment Derivation Results

The deformity assessment results are shown in Figs. 8 and 11. Overall, our ensemble approach outperformed five individual models and is comparable to the best model (ShuffleNet). The estimation results for the tibia (Fig. 9, left) show that patients with SD tend to have more deformed tibia than the non-SD, in both the legs. The NSAs in the SD groups are also more spread out (Fig. 9, right), which is consistent with the knowledge that both coxa valga (NSA >135°) and coxa vara (NSA <120°) are more frequent in patients with SD [33]. The representative output in Fig. 11 also shows how the major lower limb alignment metrics were derived. The zoomed-in view in Fig. 11 shows the stabilization of individual model key-point detection results after our ensemble approach. In general, we see PCA (green lines in femoral neck of Fig. 11) as more faithful representation of the neck axis, than other methods (RANSAC or linear regression). A

further comprehensive development into smarter and more robust ways of assessing NSA and other metrics by our downsampling technique may be warranted.

IV. Conclusion

The enLLASD has three major novelties. First, we adopted an ensemble approach to enhance robustness and reduce false positives or negatives. Second, we integrated the key-point and segmentation results for deformity assessment. Third, we derived a downsampling-based technique for major alignment metric assessments. One drawback of our approach is it may be slower to train, given the multiple models involved. But this is only linearly depending on the numbers of models, and once the models are trained, they can be deployed for fast predictions. As such, the number of models can also be scaled where necessary. Further developments may include methods to handle postoperative images, wherein implants are common.

Acknowledgment

The authors thank Dr. Cheng Chen and Dr. Grace Teng Zhang of HKU, M.D.s Tao Li, Shijie Yin, and Yapeng Zhou of HKU-SZH for helpful discussions, and Xianyou Cai (Hian-Yew Chua) of Swatow Yunyang Tech for technical support. The computations were performed on HKU-SZH Core Facility "DeepBay," a GPU cluster developed and managed by AIBD Lab.

References

[1] S. Unger et al., "Nosology of genetic skeletal disorders: 2023 revision," *Amer. J. Med. Genet. A*, vol. 191, no. 5, pp. 1164–1209, May 2023.

[2] T.-J. Cho, K. Lee, C.-W. Oh, M. S. Park, W. J. Yoo, and I. H. Choi, "Locking plate placement with unicortical screw fixation adjunctive to intramedullary rodding in long bones of patients with osteogenesis imperfecta," *J. Bone Joint Surgery-American Volume*, vol. 97, no. 9, pp. 733–737, May 2015.

[3] P. Chen, X. Zhou, H. Cai, J. Wong, Y. Hu, and M. K.-T. To, "ORCA: An ensemble deep learning framework for automatic detection and deformity assessment for lower-limb radiographs of skeletal dysplasia," in *Proc. IEEE Int. Conf. Comput. Intell. Virtual Environments Meas. Syst. Appl. (CIVEMSA)*, Jun. 2024, pp. 1–5.

[4] D. Paley, *Principles of Deformity Correction*. Berlin, Germany: Springer, 2002.

[5] C. Lindner et al., "Fully automatic segmentation of the proximal femur using random forest regression voting," *IEEE Trans. Med. Imag.*, vol. 32, no. 8, pp. 1462–1472, Aug. 2013.

[6] A. Tsai, "A deep learning approach to automatically quantify lower extremity alignment in children," *Skeletal Radiol.*, vol. 51, no. 2, pp. 381–390, Feb. 2022.

[7] Y. Pei et al., "Automated measurement of hip–knee–ankle angle on the unilateral lower limb X-rays using deep learning," *Phys. Eng. Sci. Med.*, vol. 44, no. 1, pp. 53–62, Mar. 2021.

[8] H. Archer et al., "Deep learning generated lower extremity radiographic measurements are adequate for quick assessment of knee angular alignment and leg length determination," *Skeletal Radiol.*, vol. 53, no. 5, pp. 923–933, May 2024.

[9] S. A. Sanchez, P. van Overschelde, and J. Vandemeulebroucke, "Segmentation-guided coordinate regression for robust landmark detection on X-rays: Application to automated assessment of lower limb alignment," *IEEE Access*, vol. 12, pp. 61484–61497, 2024.

[10] W. Shen et al., "Automatic segmentation of the femur and tibia bones from X-ray images based on pure dilated residual U-Net," *Inverse Problems Imag.*, vol. 15, no. 6, pp. 1333–1346, 2021.

[11] M. Hollensteiner, A. Traweger, and P. Augat, "Anatomic variability of the human femur and its implications for the use of artificial bones in biomechanical testing," *Biomed. Eng./Biomed. Tech.*, vol. 69, no. 6, pp. 551–562, Jul. 2024, doi: 10.1515/bmt-2024-0158.

[12] T. G. Dietterich, "Ensemble methods in machine learning," in *Multiple Classifier Systems*. Berlin, Germany: Springer, 2000, pp. 1–15.

[13] A. Handa, G. Grigelioniene, and G. Nishimura, "Skeletal dysplasia families: A stepwise approach to diagnosis," *RadioGraphics*, vol. 43, no. 5, May 2023, Art. no. e220067.

[14] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, Jun. 1981.

[15] K. Wada, "Labelme: Image polygonal annotation with Python," Mujin, Suwanee, GA, USA, Tech. Rep., 2016.

[16] A. Kirillov et al., "Segment anything," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2023, pp. 4015–4026.

[17] Ultralytics.(2023). *YOLOv8*. [Online]. Available: https://github.com/ultralytics/ultralytics

[18] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2016, pp. 483–499.

[19] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An extremely efficient convolutional neural network for mobile devices," Megvii, Beijing, China, Tech. Rep., 2018.

[20] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 5686–5696.

[21] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4510–4520.

[22] T. Jiang et al., "RTMPose: Real-time multi-person pose estimation based on MMPose," 2023, *arXiv:2303.07399*.

[23] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[24] R. P. K. Poudel, S. Liwicki, and R. Cipolla, "Fast-SCNN: Fast semantic segmentation network," 2019, *arXiv:1902.04502*.

[25] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.

[26] W. Zhang, J. Pang, K. Chen, and C. Change Loy, "K-Net: Towards unified image segmentation," 2021, *arXiv:2106.14855*.

[27] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR*, Jul. 2016, pp. 6230–6239.

[28] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," 2015, *arXiv:1505.04597*.

[29] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Álvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," in *Proc. Neural Inf. Process. Syst.*, Jan. 2021, pp. 12077–12090.

[30] N. Ravi et al., "SAM 2: Segment anything in images and videos," 2024, *arXiv:2408.00714*.

[31] J. Ma, Y. He, F. Li, L. Han, C. You, and B. Wang, "Segment anything in medical images," *Nature Commun.*, vol. 15, no. 1, p. 654, 2024.

[32] K. Chen et al.. (2020). *OpenMMLab: A Comprehensive Computer Vision Research and Deployment Platform*. [Online]. Available: https://github.com/open-mmlab

[33] C.-W. Oh, M. M. Thacker, W. G. Mackenzie, and E. C. Riddle, "Coxa vara: A novel measurement technique in skeletal dysplasias," *Clin. Orthopaedics Rel. Res.*, vol. 447, pp. 125–131, Jun. 2006.

Peikai Chen received the B.Eng. degree from Zhejiang University (ZJU), Hangzhou, China in 2006, and the Ph.D. degree from HKU, Hong Kong, in 2012.

He also attended Class ACEE of Chu Ko Chen Honors College at ZJU. He is currently a Research Scientist and the Executive Head of the AIBD Laboratory at HKU-SZH, Shenzhen, China, with a focus on biomedical engineering and biomedical informatics.

Dr. Chen won a national second prize in the National Undergraduate Electronic Design Contest in 2005, the HKU UPF award in 2007, and the Lindon Eaves Best Poster Award (Boulder, CO, USA) in 2015.

Xinlin Zhou received the master's degree in engineering from Shenzhen University, Shenzhen, China, in 2024.

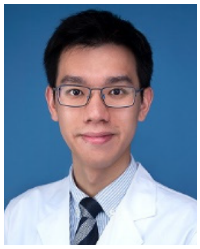
He was a Research Assistant at HKU-SZH, Shenzhen. He is currently working as a Physics Teacher with High School, Guangzhou.

Haihua Cai received the bachelor's degree in mechanical engineering from Shantou University, Shantou, China, in 2023. He is currently pursuing the master's degree in engineering with Shenzhen University, Shenzhen, China.

He is currently a Research Assistant at HKU-SZH, Shenzhen.

David J. H. Shih received the B.Sc., M.Sc., and Ph.D. degrees from the University of Toronto, Toronto, ON, Canada, in 2008, 2011, and 2015, respectively.

He completed post-doctoral training at Dana-Farber, Boston, USA, and MD Anderson, Houston, USA. He was a Research Assistant Professor at UTHealth, Houston, from 2020 to 2022. He is currently an Assistant Professor at HKU, and focuses on cancer genomics, tumor evolution, and DNA repair using multiomics and EHR data.



Janus S. H. Wong received the Medical degree from HKU, Hong Kong, in 2015.

He completed orthopedic residency at major hospitals, including Queen Mary, Hong Kong, and Prince of Wales, Randwick, NSW, Australia. He joined HKU as a Clinical Assistant Professor in 2019 and completed higher training in 2022.

Yong Hu (Senior Member, IEEE) received the B.Sc. and M.Sc. degrees from Tianjin University, Tianjin, China, in 1985 and 1988, respectively, and the Ph.D. degree from HKU, Hong Kong, in 1999.

He is an Associate Professor at the Department of Orthopedics and Traumatology, HKU, the Head of the AIBD Lab at HKU-SZH, and the Director of the Laboratory of Neural Engineering (Shenzhen). His research focuses on neural engineering, clinical electrophysiology, and biomedical signal processing.

Michael Kai-Tsun To is a Clinical Professor of pediatric orthopedics at HKU, Hong Kong. Clinically, he focuses on metabolic bone disease, neuromuscular disorders, and pediatric trauma. He has published widely in top journals and co-authored academic books. His research interests include nanomedicine for tissue regeneration, microfluidic drug delivery, and genetics of skeletal dysplasia. He holds fellowships, including FRCS(Edin), FHKCOS, and FHKAM.